31759-190645
Akihiro OKUMURA et al.
filed 6/27/03

日 本 国 特 許 庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出願年月日
Date of Application: ２００２年 ６月２７日

出願番号
Application Number: 特願２００２－１８７６２２

[ST.10/C]: [ＪＰ２００２－１８７６２２]

出願人
Applicant(s): 沖電気工業株式会社

２００３年 ４月 １日

特許庁長官
Commissioner,
Japan Patent Office 太田信一郎

出証番号 出証特２００３－３０２２４４０

【書類名】　特許願
【整理番号】　KN002520
【提出日】　平成14年 6月27日
【あて先】　特許庁長官　及川　耕造　殿
【国際特許分類】　G06F 15/00
【発明者】
　【住所又は居所】　東京都港区虎ノ門１丁目７番１２号　沖電気工業株式会社内
　【氏名】　奥村　晃弘
【発明者】
　【住所又は居所】　東京都港区虎ノ門１丁目７番１２号　沖電気工業株式会社内
　【氏名】　池野　篤司
【特許出願人】
　【識別番号】　000000295
　【氏名又は名称】　沖電気工業株式会社
　【代表者】　篠塚　勝正
【代理人】
　【識別番号】　100090620
　【弁理士】
　【氏名又は名称】　工藤　宣幸
【手数料の表示】
　【予納台帳番号】　013664
　【納付金額】　21,000円
【提出物件の目録】
　【物件名】　明細書　1
　【物件名】　図面　1
　【物件名】　要約書　1
　【包括委任状番号】　9006358

【プルーフの要否】　　要

【書類名】　　明細書

【発明の名称】　主要領域判定装置

【特許請求の範囲】

【請求項１】　構造化文書から主要な領域を判定する装置において、

定期的または不定期に構造化文書を取得する読み込み部と、

読み込んだ構造化文書を１つ以上の領域に分割する分割部と、

分割結果を一時的に記憶する分割結果記憶部と、

領域ごとに前回の内容と今回の内容を比較して更新有無を調べる比較部と、

領域ごとに過去の更新情報を記憶する更新情報記憶部と、

前回までの更新情報と今回の更新有無から更新頻度を算出する更新頻度算出部と、

更新頻度が最も高い領域を主要領域として判定する判定部と

を備えることを特徴する主要領域判定装置。

【請求項２】　構造化文書から主要な領域を判定する装置において、

定期的または不定期に構造化文書を取得する読み込み部と、

読み込んだ構造化文書を１つ以上の領域に分割する分割部と、

読み込み結果を一時的に記憶する記憶部と、

領域ごとに前回の内容と今回の内容を比較して更新有無を調べる比較部と、

領域ごとに過去の更新情報を記憶する更新情報記憶部と、

前回までの更新情報と今回の更新有無から更新頻度を算出する更新頻度算出部と、

更新頻度が最も高い領域を主要領域として判定する判定部と

を備えることを特徴する主要領域判定装置。

【請求項３】　構造化文書から主要な領域を判定する装置において、

定期的または不定期に構造化文書を取得する読み込み部と、

読み込んだ構造化文書を１つ以上の領域に分割する分割部と、

分割した領域の内容を対応するシンボルに変換する変換部と、

各領域の内容を変換して求めたシンボルを一時的に記憶する記憶部と、

領域ごとに前回のシンボルと今回のシンボルを比較して更新有無を調べる比較部と、

領域ごとの更新情報を記憶する更新情報記憶部と、

前回までの更新情報と今回の更新有無から更新頻度を算出する更新頻度算出部と、

更新頻度が最も高い領域を主要領域として判定する判定部と

を備えることを特徴する主要領域判定装置。

【請求項４】　前記更新情報に更新頻度の値を用い、過去の更新頻度の値と今回の更新有無から決定した値との指数平均を用いて新たな更新頻度を算出することを特徴とする請求項１～３のいずれかに記載の主要領域判定装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、構造化文書の記述から主要な領域を判定し出力する主要領域判定装置に関し、例えば、WWW (World Wide Web) サイトから構造化文書を取得するシステムに適用し得るものである。

【０００２】

【従来の技術】

WWWサイトに存在する構造化文書を取得し閲覧するためのツールとして、WWWブラウザがある。一般的に、構造化文書はその文書のページのレイアウト、文字の大きさなどを柔軟に指定することができるようになっている。特に、図１のように、タイトル（領域Ａ）、他の構造化文書へのリンク（領域Ｂ）、本文（領域Ｃ）、など、ページがいくつかの領域に分割されて、WWWブラウザに表示されるような構造化文書が多く見られる。WWWブラウザを用いて、このような構造化文書から必要な情報を得るためには、ユーザは、目的の構造化文書のＵＲＬを指定し、その文書がWWWブラウザ上に表示された後に、文書をスクロールしながら目視により検索したり（人手による検索）、あるいは文字列検索機能を利用するといった作業を行なう必要がある。例えば、図１の領域Ｃが、ユーザの必要とする文書であったとし、こういった構造化文書が多数ある場合には、その

ユーザが必要とする情報のみを複数の構造化文書から自動的にスクラップし、１つの文書にまとめてユーザに提示されることが、人手による作業を簡略化する上で望ましくなる。このようなＷＷＷ情報抽出システムが、特開平１０－１８７７５３号公報に示されている。

【０００３】

【発明が解決しようとする課題】

しかしながら、上記におけるＷＷＷ情報抽出システムでは、ユーザが構造化文書中で自分が必要とするデータの開始箇所と終了箇所をあらかじめ手入力により指定することが必要である。このため、大量の構造化文書に対して実施するにはユーザの負担が大きく現実的ではなかった。

【０００４】

本発明はこれらの問題点を解決するためになされたものであり、構造化文書における主要領域を判定することを目的とする。

【０００５】

【課題を解決するための手段】

図１の例からも分かるように、主要領域以外の領域にはタイトル（領域Ａ）や他の構造化文書へのリンク（領域Ｂ）などが存在する。これらは主要領域の記述と比較して更新頻度が低いという特徴がある。そこで、本発明では領域毎の更新頻度に基づいて主要領域を判定するように構成する。

【０００６】

第１の本発明は、構造化文書から主要な領域を判定する主要領域判定装置において、定期的または不定期に構造化文書を取得する読み込み部と、読み込んだ構造化文書を１つ以上の領域に分割する分割部と、分割結果を一時的に記憶する分割結果記憶部と、領域ごとに前回の内容と今回の内容を比較して更新有無を調べる比較部と、領域ごとに過去の更新情報を記憶する更新情報記憶部と、前回までの更新情報と今回の更新有無から更新頻度を算出する更新頻度算出部と、更新頻度が最も高い領域を主要領域として判定する判定部とを備えることを特徴する。

【０００７】

第２の本発明は、構造化文書から主要な領域を判定する主要領域判定装置にお

いて、定期的または不定期に構造化文書を取得する読み込み部と、読み込んだ構造化文書を１つ以上の領域に分割する分割部と、読み込み結果を一時的に記憶する記憶部と、領域ごとに前回の内容と今回の内容を比較して更新有無を調べる比較部と、領域ごとに過去の更新情報を記憶する更新情報記憶部と、前回までの更新情報と今回の更新有無から更新頻度を算出する更新頻度算出部と、更新頻度が最も高い領域を主要領域として判定する判定部とを備えることを特徴する。

【０００８】

第３の本発明は、構造化文書から主要な領域を判定する主要領域判定装置において、定期的または不定期に構造化文書を取得する読み込み部と、読み込んだ構造化文書を１つ以上の領域に分割する分割部と、分割した領域の内容を対応するシンボルに変換する変換部と、各領域の内容を変換して求めたシンボルを一時的に記憶する記憶部と、領域ごとに前回のシンボルと今回のシンボルを比較して更新有無を調べる比較部と、領域ごとの更新情報を記憶する更新情報記憶部と、前回までの更新情報と今回の更新有無から更新頻度を算出する更新頻度算出部と、更新頻度が最も高い領域を主要領域として判定する判定部とを備えることを特徴する。

【０００９】

【発明の実施の形態】

（Ａ）第１の実施形態

（Ａ－１）第１の実施形態の構成

図２は、第1の実施形態の主要領域判定装置の構成を示すブロック図である。第1の実施形態の主要領域判定装置は、パソコン等の通信機能を有する情報処理装置上で実現されるが、機能的には、読み込み部１０１、バッファ部１０２、分割部１０３、分割結果記憶部１０４、領域内容比較部１０５、更新頻度算出部１０６、更新頻度記憶部１０７及び判定部１０８を有する。

【００１０】

読み込み部１０１は、定期的にあるいは不定期（利用者が読み込みを指示した時など）に指定された位置の構造化文書の内容をバッファ部１０２に読み込む機能部である。

【００１１】

バッファ部１０２は、読み込み部１０１が読み込んだ内容や分割部１０３が処理した結果を一時的に記憶する機能部である。

【００１２】

分割部１０３は、バッファ部１０２に読み込まれた構造化文書を解析することによって、予め指定された文書構造に基づいて、バッファ部１０２の内容を領域に分割する機能部である。

【００１３】

分割結果記憶部１０４は、前回に分割部１０３が分割した結果と今回分割した結果を比較するために、前回の分割結果を記憶しておく機能部である。

【００１４】

領域内容比較部１０５は、分割結果記憶部１０４が記憶している前回に読み込んだ内容の分割結果とバッファ部１０２が記憶している今回読み込んだ内容の分割結果とを領域ごとに内容を比較して更新の有無を検出する機能部である。

【００１５】

更新頻度算出部１０６は、領域内容比較部１０５が検出した更新有無と、前回までの更新頻度から今回の更新頻度を算出する機能部である。

【００１６】

更新頻度記憶部１０７は、分割領域ごとに内容が更新される頻度を更新頻度Ｓとして記憶する機能部である。

【００１７】

判定部１０８は、更新頻度記憶部１０７の記憶する更新頻度に基づいて最も頻繁に内容が変化している領域を主要領域と判定し、バッファ部１０２から該当領域の内容を取り出してそれを出力する機能部である。

【００１８】

（Ａ－２）第１の実施形態の動作

第１の実施形態の主要領域判定装置は、定期的にあるいは不定期（利用者が読み込みを指示した時など）に構造化文書の内容を前回と比較することにより、主要領域を判断するための情報を更新する。このため、「情報更新」時と、「主要

領域の内容出力」時の２つの動作がある。

【００１９】

図３は、第１の実施形態の「情報更新」時の動作を示すフローチャートである。以下、図３のフローチャートについて説明する。

【００２０】

まず、ステップＳ１０１で予め指定された位置の構造化文書の内容をバッファ部１０２に読み込む。このときに、他の構造化文書の挿入が必要な場合はそれを実施する。図４の構成の場合に、ｉｎｄｅｘ．ｈｔｍｌを指定されたときの読み込み結果の例を図５に示す。

【００２１】

次に、ステップＳ１０２でバッファ部１０２に読み込まれた構造化文書を予め定められた文書構造に基づいて分割し、分割したそれぞれの領域は文書の最初に表れたものから順に１から始まる領域番号を付与する。例えば、図５の内容をフレーム単位で分割する場合、処理結果（バッファ部１０２に書き込まれる内容）は図６のようになる。

【００２２】

ステップＳ１０３では分割結果記憶部１０４が記憶している前回に読み込んだ構造化文書を分割した結果と、バッファ部１０２が記憶している今回読み込んだ構造化文書を分割部１０３が分割した結果とを比較して、領域ごとに更新頻度Ｓを算出し、更新頻度記憶部１０７の内容を更新する。更新頻度Ｓの算出方法の詳細については後述する。

【００２３】

ステップＳ１０４では、バッファ部１０２が記憶している内容を上書きすることによって、分割結果記憶部１０４の内容を更新する。この内容は次回の「情報更新」時に参照される。

【００２４】

以上が、「情報更新」時の動作である。

【００２５】

図７は、第１の実施形態の「主要領域の内容出力」時の動作を示すフローチャ

ートである。「情報更新」時のフローチャートと同じ部分は同一のステップ番号を付与し重複して説明することは省略する。

【００２６】

以下、図７のフローチャートについて説明する。

【００２７】

ステップＳ１０１およびステップＳ１０２は「情報更新」時のフローチャートと同じである。

【００２８】

ステップＳ１０５では、更新頻度記憶部１０７から更新頻度Ｓの値が最も小さい領域番号を得る。このとき、複数の候補がある場合は領域番号の値が小さいものを優先させることにより、１つの領域番号を決定する。そして、バッファ部１０２から該当領域番号の内容を取り出して、これを主要領域の内容として出力し、処理を終了する。

【００２９】

図８は、ステップ１０３で更新頻度Ｓを算出する方法の詳細について説明したものである。

【００３０】

以下、図８のフローチャートについて説明する。

【００３１】

ステップＳ１５１で、バッファ部１０２内の分割された領域から１つずつ領域を選択して、ステップＳ１５２からステップＳ１５９までの処理を繰り返す。

【００３２】

ステップＳ１５２では、領域内容比較部１０５がステップＳ１５１で選択した領域の内容と、分割結果記憶部１０４内の同じ領域番号が示す領域の内容同士を比較する。

【００３３】

ステップＳ１５３で内容を比較した結果、同じだった場合はステップＳ１５５へ、違った場合はステップＳ１５４へ進む。比較する対象が存在しない場合は、比較結果が違った場合と同様にステップＳ１５４へ進む。

【００３４】

ステップＳ１５４では、バッファ部１０２での分割数と分割結果記憶部１０４での分割数を比較する。分割数が同じ場合はステップＳ１５６へ、違う場合はステップＳ１５８へ進む。

【００３５】

ステップＳ１５５では、前回の内容と同じなので得点Ｐに１００点を設定する。

【００３６】

ステップＳ１５６では、前回と内容が異なっているので得点Ｐに０点を設定する。

【００３７】

ステップＳ１５７では、指数平均によって更新頻度Ｓの値を求める。具体的には、更新頻度記憶部１０７から該当領域の更新頻度を取得し、これを前回の更新頻度Ｓ０とし、この値と得点Ｐの値を使ってＳ＝Ｓ０・α＋Ｐ（１－α） を計算して求める。αは０＜α＜１の定数で予め適当な値を決定しておく。

【００３８】

ステップＳ１５８では、前回とは分割数も内容も違っているため、前回までの更新頻度との関連が不明である。このため、更新頻度Ｓに０を設定し初期化する。

【００３９】

ステップＳ１５９では、これまでのステップで求めた更新頻度Ｓの値を更新頻度記憶部１０７の該当領域の値として更新する。

【００４０】

ステップＳ１６０では、バッファ部１０２内の分割された領域全てに対して処理が終わるまで繰り返す。

【００４１】

以上が更新頻度Ｓを算出する方法である。

【００４２】

このような計算をすることにより、更新頻度Ｓは前回と内容が同じことが多い

領域ほど大きな値をとるようになり、内容が更新される頻度を表すようになる。つまり、値が大きいほど内容が更新される頻度が低く、値が小さいほど内容の更新が頻繁であることを示す。

【００４３】

次に、図９および図１０を使って更新頻度Ｓの算出の具体例を説明する。

【００４４】

図９は分割結果記憶部１０４の分割数と、バッファ部１０２の分割数が共に３であり、領域３の内容のみ違いがある場合である。領域１，２，３の前回の更新頻度Ｓ０をそれぞれ７３，７３，４６とし、α＝０．８とすると、更新頻度Ｓの値は以下の計算より、それぞれ７８，７８，３７となる。

【００４５】

７３×０．８＋１００×（１－０．８）≒７８

４６×０．８＋０×（１－０．８）≒３７

図１０は分割結果記憶部１０４の分割数が３で、バッファ部１０２の分割数が４であり、領域１の内容のみが同じである場合である。領域１，２，３の前回の更新頻度Ｓ０をそれぞれ７３，７３，４６とし、α＝０．８とすると、内容が同じ領域１の更新頻度Ｓの値は上記の計算により７８となる。内容が同じではない領域２～４は、更新頻度Ｓの値が０に初期化される。

【００４６】

以上が第１の実施形態の動作である。

【００４７】

「情報更新」動作を適宜実行しておくことにより、更新頻度記憶部１０７の内容が最適化され、「主要領域の内容出力」時に正しく主要領域を出力することができるようになる。

【００４８】

ここまでで、「情報更新」動作と「主要領域の内容出力」動作の２種類の動作について説明したが、「主要領域の内容出力」時に同時に「情報更新」を実行しても良い。この場合、２つの動作を組み合わせて、フローチャートは図１１のようになる。

【００４９】

（Ａ－３）第１実施形態の効果

第１実施形態によれば、以下の効果を奏することができる。

【００５０】

自然言語処理を用いないので、記述言語に依存せずに主要領域を判定することができる。

【００５１】

構造化文書の解析を実施するが、予め指定した文書構造だけを処理すればよいので、全ての解析をするよりも処理が非常に軽い。

【００５２】

自動的に主要領域を判定できるので、(i)指定ウェブページの更新時の通知（主要領域以外の更新は通知しない、など）、(ii)検索（主要領域以外は検索対象としない、など）、(iii)要約（主要領域のみを要約対象とする、など）などのサービスやシステムを容易に構築することが可能となる。

【００５３】

分割結果記憶部１０４では分割した結果をそのまま記憶するので、後述する第２の実施形態のようにチェックサムを算出する必要がなく、計算量が少ない。

【００５４】

（Ｂ）第２の実施形態

（Ｂ－１）第２の実施形態の構成

図１２は、第２の実施形態の主要領域判定装置の機能的構成を示すブロック図である。

【００５５】

第１の実施形態では、分割結果の内容をそのまま記憶したが、内容から算出した値（チェックサム）を記憶するように構成した。第２実施形態の構成を以下で説明する。但し、第１実施形態と同じ構成要素については、同一番号を付して繰り返し説明することは省略する。

【００５６】

第２の実施形態では、第１の実施形態の分割結果記憶部１０４及び領域内容比

較部１０５に代え、チェックサム算出部２０１、チェックサム記憶部２０２及びチェックサム比較部２０３が設けられている。

【００５７】

チェックサム算出部２０１は、分割結果の内容からチェックサムを算出する機能部である。

【００５８】

チェックサム記憶部２０２は、各分割結果から算出したチェックサムの値を記憶しておく機能部である。

【００５９】

チェックサム比較部２０３は、チェックサム記憶部２０２が記憶している前回のチェックサムとバッファ部１０２が記憶している今回のチェックサムとを領域ごとに値が同じかどうか比較する機能部である。

【００６０】

（Ｂ－２）第２の実施形態の動作

第１の実施形態と同様に、「情報更新」時と「主要領域の内容出力」時の２つの動作があるが、「主要領域の内容出力」時の動作は第１実施形態と同じなので省略し、「情報更新」時の動作についてのみ説明する。

【００６１】

図１３は、第２の実施形態の「情報更新」時の動作を示すフローチャートである。第1の実施形態と同じ部分は、同一ステップ番号を付与し重複して説明することは省略する。

【００６２】

以下、図１３のフローチャートについて説明する。

【００６３】

ステップＳ１０１およびステップＳ１０２は第１の実施形態と同じである。

【００６４】

ステップＳ２０１では、チェックサム記憶部２０２が記憶している前回の各領域のチェックサムと、バッファ部１０２が記憶している今回読み込んだ構造化文書の各領域の内容からチェックサム算出部２０１が算出したチェックサムとを比

較して、領域ごとに更新頻度Ｓを算出し、更新頻度記憶部１０７の内容を更新する。更新頻度Ｓの算出方法の詳細については後述する。

【００６５】

ステップＳ２０２では、バッファ部１０２が記憶している分割結果からチェックサム算出部２０１がそれぞれの領域のチェックサムを算出し、チェックサム記憶部２０２の内容を更新する。

【００６６】

チェックサムの算出方法としては、分割領域の内容をバイトデータ列とみなして、全てのデータを加算した結果の下位４バイトを取るなどとすればよい。

【００６７】

この内容は次回の「情報更新」時に参照される。

【００６８】

以上が、「情報更新」時の動作である。

【００６９】

図１４は、ステップＳ２０１で更新頻度Ｓを算出する方法の詳細について説明したものである。

【００７０】

第1の実施形態と同じ部分は、同一ステップ番号を付与し重複して説明することは省略する。

【００７１】

以下、図１４のフローチャートについて説明する。

【００７２】

ステップＳ２５１で、バッファ部１０２内の分割された領域から１つづつ領域を選択して、ステップＳ２５２からステップＳ１５９までの処理を繰り返す。

【００７３】

ステップＳ２５２では、ステップＳ２５１で選択した領域の内容からチェックサム算出部２０１が算出したチェックサムと、チェックサム記憶部２０２内の同じ領域番号が示す領域のチェックサム同士を比較する。

【００７４】

ステップＳ２５２でチェックサムを比較した結果、同じだった場合はステップＳ１５５へ、違った場合はステップＳ１５４へ進む。

【００７５】

ステップＳ１５４～ステップＳ１６０は第１実施形態と同じである。

【００７６】

以上が更新頻度Ｓを算出する方法である。

【００７７】

チェックサム同士を比較するところが第１実施形態との違いである。

【００７８】

以上が第２の実施形態の動作である。

【００７９】

（Ｂ－３）第２の実施形態の効果

第２実施形態によっても、（１）自然言語処理を用いないので、記述言語に依存せずに主要領域を判定することができる、（２）構造化文書の解析を実施するが、予め指定した文書構造だけを処理すればよいので、全ての解析をするよりも処理が非常に軽い、（３）自動的に主要領域を判定できるので、(i)指定ウェブページの更新時の通知（主要領域以外の更新は通知しない、など）、(ii)検索（主要領域以外は検索対象としない、など）、(iii)要約（主要領域のみを要約対象とする、など）などのサービスやシステムを容易に構築することが可能となる、等の効果を奏する。

【００８０】

また、第２の実施形態によれば、チェックサム記憶部２０２では各領域の内容から算出したチェックサムを記憶するので、第１実施形態のように、分割結果をそのまま記憶する場合と比較して、記憶容量が小さくてすむ、という効果をも奏する。

【００８１】

（Ｃ）他の実施形態

第１の実施形態で、前回との比較を実施するために分割した結果を記憶しておくように構成したが、読み込んだ内容をそのまま記憶しておき、比較する前に改

めて分割するように構成してもよい。

【００８２】

第１および第２の実施形態で、分割した領域に対して文書の始めから表れた順序に従って対応付けを実施したが、構造化文書の内部でそれぞれの領域を識別することができる情報が付加されている場合は、その情報を使うように構成してもよい。

【００８３】

第１および第２の実施形態で、ＨＴＭＬを例に説明したが、ＸＭＬやＳＧＭＬなどでもよい。

【００８４】

第２の実施形態でチェックサムを使うと説明したが、ハッシュ関数から求めたハッシュ値を使うようにしてもよい。

【００８５】

１つの構造化文書を処理する場合について説明したが、複数の構造化文書を対象としてもよい。この場合、分割結果記憶部１０４、更新頻度記憶部１０７、チェックサム記憶部２０２は構造化文書ごとに独立に情報を記憶する。

【００８６】

第１および第２の実施形態では、構造化文書がWWWサイト上に公開されたものであることを前提としているが、記録媒体から得た構造化文書等、対象とする構造化文書の入手方法は問われないものである。

【００８７】

【発明の効果】

本発明の主要領域判定装置によれば、構造化文書における主要領域を判定することができる。

【図面の簡単な説明】

【図１】

構造化文書の説明図である。

【図２】

第１の実施形態の主要領域判定装置の機能的構成を示すブロック図である。

【図３】

第１の実施形態の「情報更新」時の動作を示すフローチャートである。

【図４】

第１の実施形態の動作説明に用いる構造化文書の例を示す説明図である。

【図５】

図４の構造化文書の所定ファイルを読み込んだ結果例を示す説明図である。

【図６】

図５の構造化文書の領域の説明図である。

【図７】

第１の実施形態の「主要領域の内容出力」時の動作を示すフローチャートである。

【図８】

第１の実施形態の更新頻度Ｓの算出方法を示すフローチャートである。

【図９】

第１の実施形態の更新頻度Ｓの算出の具体例の説明図（１）である。

【図１０】

第１の実施形態の更新頻度Ｓの算出の具体例の説明図（２）である。

【図１１】

第１の実施形態の「情報更新」時の動作と「主要領域の内容出力」時の動作とまとめた場合のフローチャートである。

【図１２】

第２の実施形態の主要領域判定装置の機能的構成を示すブロック図である。

【図１３】

第２の実施形態の「情報更新」時の動作を示すフローチャートである。

【図１４】

第２の実施形態の更新頻度Ｓの算出方法を示すフローチャートである。

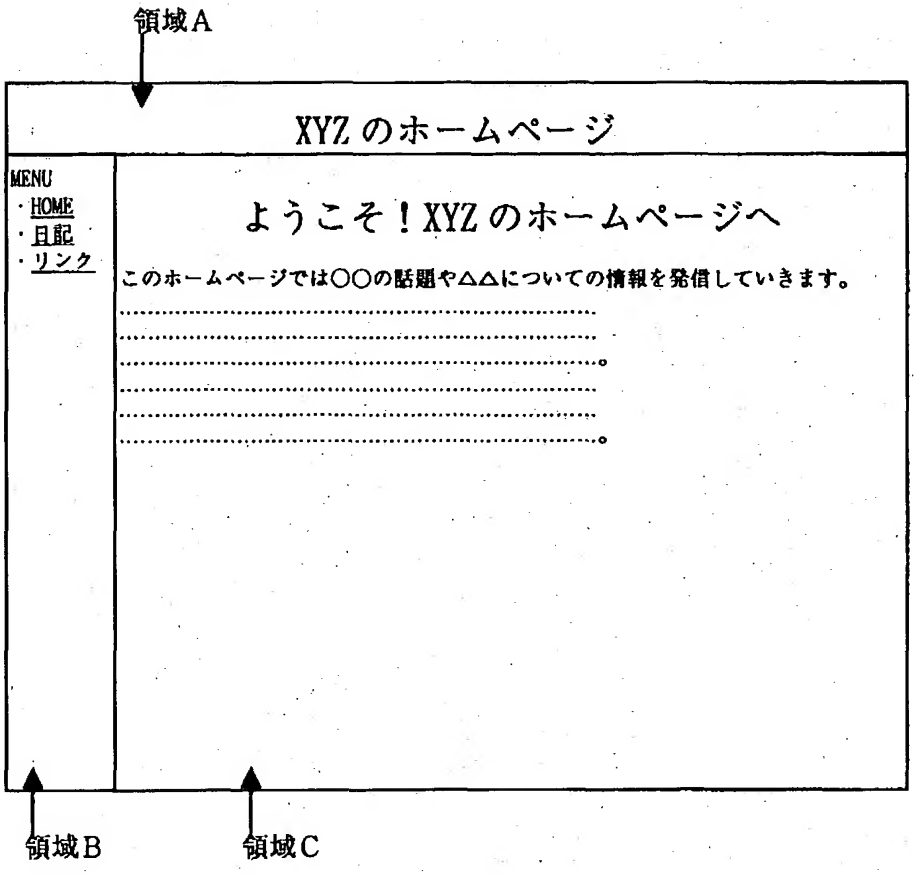
【符号の説明】

１０１…読み込み部、１０２…バッファ部、１０３…分割部、１０４…分割結果記憶部、１０５…領域内容比較部、１０６…更新頻度算出部、１０７…更新頻

度記憶部、１０８…判定部、２０１…チェックサム算出部、２０２…チェックサム記憶部、２０３…チェックサム比較部。

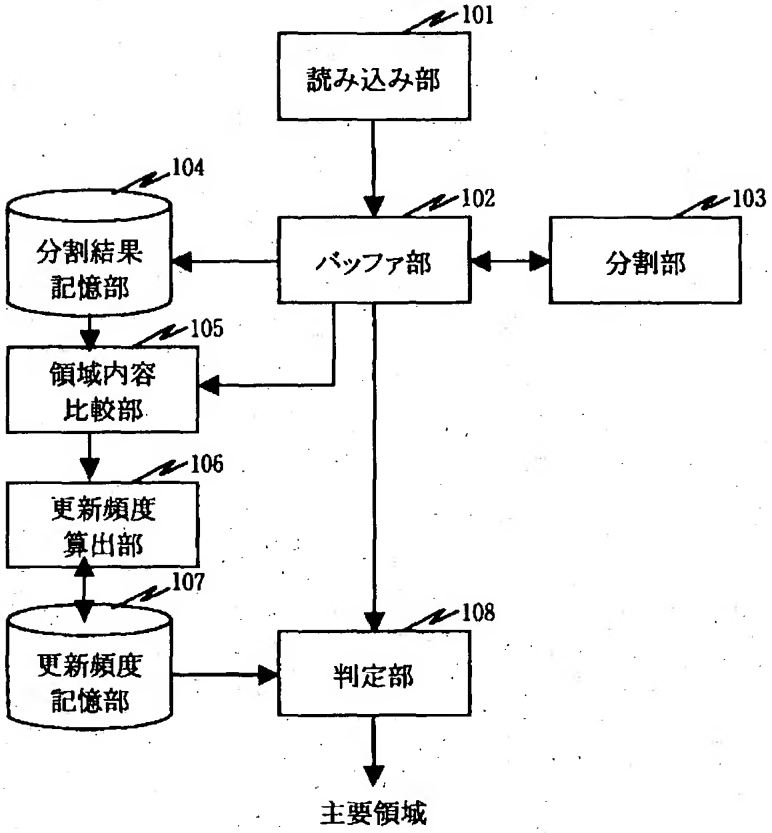
【書類名】　　図面

【図１】

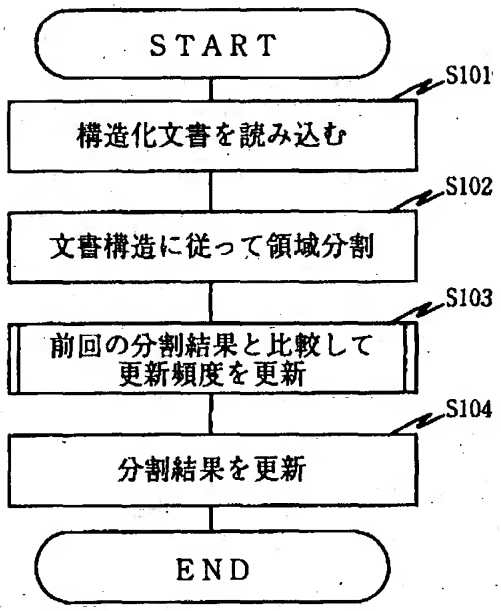


出証特２００３－３０２２４４０

【図２】


101
読み込み部
104
分割結果
記憶部
102
バッファ部
103
分割部
105
領域内容
比較部
106
更新頻度
算出部
107
更新頻度
記憶部
108
判定部
主要領域

【図３】


START
S101
構造化文書を読み込む
S102
文書構造に従って領域分割
S103
前回の分割結果と比較して
更新頻度を更新
S104
分割結果を更新
END

【図４】

index.html

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Frames//EN">
<HTML>
<HEAD>
<TITLE>XYZ のホームページ</TITLE>
</HEAD>
<FRAMESET rows="10%,*">
<FRAME src="title.html" name ="title">
<FRAMESET cols="15%,*">
<FRAME src="menu.html" name ="menu">
<FRAME src="main.html" name="main">
<NOFRAMES>
<P>フレーム非対応ブラウザをお使いの方はこちらから</P>
<P><A href="menu.htm">MENU</A>
<A href="main.htm">メイン</A></P>
</NOFRAMES>
</FRAMESET>
</HTML>
```

title.html

```
<HTML>
<HEAD>
<TITLE>タイトル</TITLE>
</HEAD>
<BASEFONT size=7 color=#008000>
<CENTER>
XYZ のホームページ
</CENTER>
</HTML>
```

menu.html

```
<HTML>
<HEAD>
<TITLE>メニュー</TITLE>
</HEAD>
<BODY>
<BASEFONT size=5 color="blue">
<ULL>MENU
<LI><A href="home.html"  target="main">HOME</A>
<LI><A href="diary.html" target="main">日記</A>
<LI><A href="links.html" target="main">リンク</A>
</ULL>
</BODY>
</HTML>
```

main.html

```
<HTML>
<HEAD>
<TITLE>本文</TITLE>
</HEAD>
<BASEFONT size=5 color="black">
<BR>
<P>
<CENTER>
<FONT size=7>
ようこそ！XYZ のホームページへ<BR>
</CENTER>
</FONT>
</P>
このホームページでは○○の話題や△△についての
情報を発信していきます。
……
……
……。
……
……
……。
</HTML>
```

【図５】

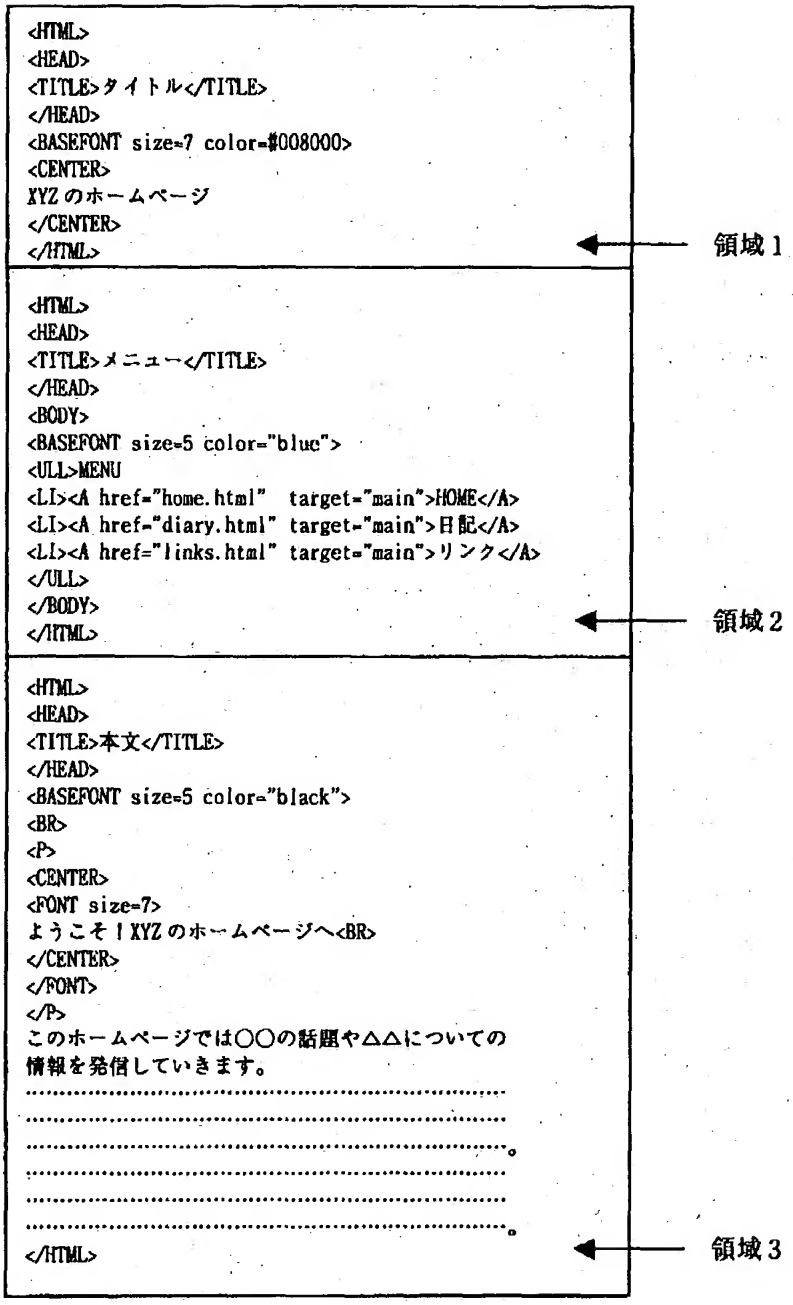
```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Frames//EN">
<HTML>
<HEAD>
<TITLE>XYZ のホームページ</TITLE>
</HEAD>
<FRAMESET rows="10%,*">
<FRAME src="title.html" name ="title">
<HTML>
<HEAD>
<TITLE>タイトル</TITLE>
</HEAD>
<BASEFONT size=7 color=#008000>
<CENTER>
XYZ のホームページ
</CENTER>
</HTML>
<FRAMESET cols="15%,*">
<FRAME src="menu.html" name ="menu">
<HTML>
<HEAD>
<TITLE>メニュー</TITLE>
</HEAD>
<BODY>
<BASEFONT size=5 color="blue">
<ULL>MENU
<LI><A href="home.html"  target="main">HOME</A>
<LI><A href="diary.html" target="main">日記</A>
<LI><A href="links.html" target="main">リンク</A>
</ULL>
</BODY>
</HTML>
<FRAME src="main.html" name="main">
<HTML>
<HEAD>
<TITLE>本文</TITLE>
</HEAD>
<BASEFONT size=5 color="black">
<BR>
<P>
<CENTER>
<FONT size=7>
ようこそ！XYZ のホームページへ<BR>
</CENTER>
</FONT>
</P>
このホームページでは○○の話題や△△についての
情報を発信していきます。
.....
.....
.....。
.....
.....
.....。
</HTML>
<NOFRAMES>
<P>フレーム非対応ブラウザをお使いの方はこちらから</P>
<P><A href="menu.htm">MENU</A>
<A href="main.htm">メイン</A></P>
</NOFRAMES>
</FRAMESET>
</HTML>
```

title.html の内容を挿入

menu.html の内容を挿入

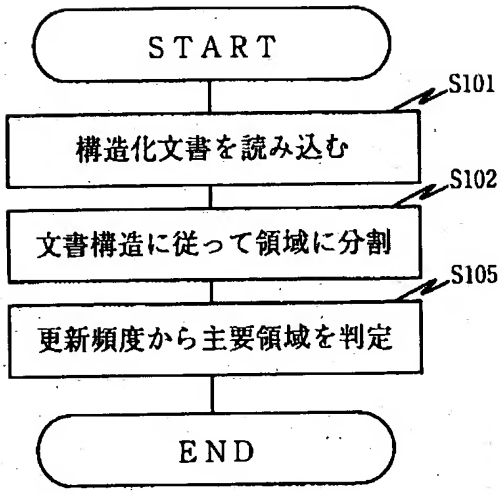
main.html の内容を挿入

【図６】

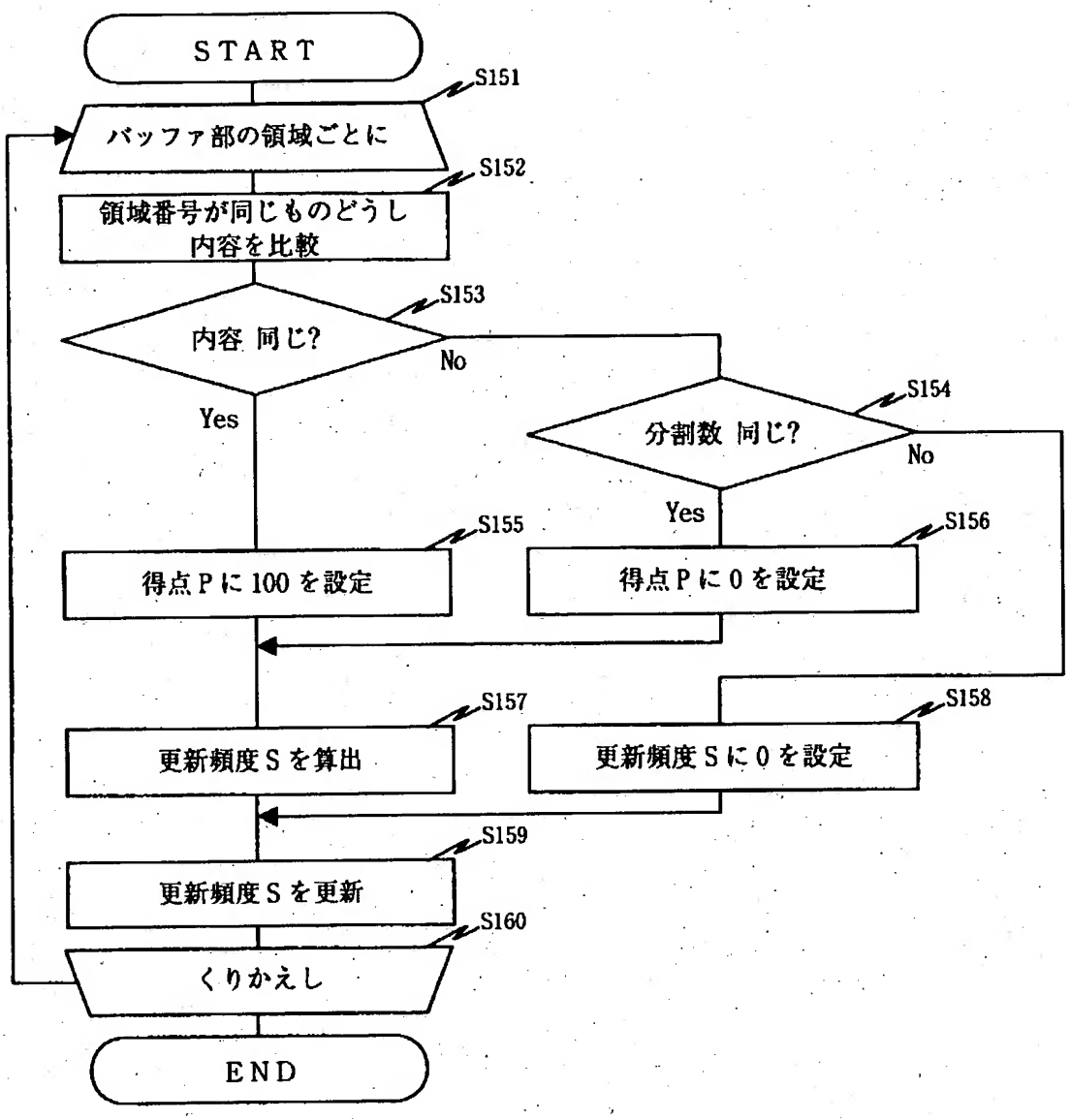

<HTML>
<HEAD>
<TITLE>タイトル</TITLE>
</HEAD>
<BASEFONT size=7 color=#008000>
<CENTER>
XYZのホームページ
</CENTER>
</HTML>
領域１
<HTML>
<HEAD>
<TITLE>メニュー</TITLE>
</HEAD>
<BODY>
<BASEFONT size=5 color="bluc">
<ULL>MENU
<LI><A href="home.html" target="main">HOME</A>
<LI><A href="diary.html" target="main">日記</A>
<LI><A href="links.html" target="main">リンク</A>
</ULL>
</BODY>
</HTML>
領域２
<HTML>
<HEAD>
<TITLE>本文</TITLE>
</HEAD>
<BASEFONT size=5 color="black">

<P>
<CENTER>
<FONT size=7>
ようこそ！XYZのホームページへ
</CENTER>
</FONT>
</P>
このホームページでは○○の話題や△△についての
情報を発信していきます。
……
……
……。
……
……
……。
</HTML>
領域３

【図７】


START
S101
構造化文書を読み込む
S102
文書構造に従って領域に分割
S105
更新頻度から主要領域を判定
END

【図８】


START
S151
バッファ部の領域ごとに
S152
領域番号が同じものどうし
内容を比較
S153
内容 同じ?
No
Yes
S154
分割数 同じ?
No
Yes
S155
得点Pに100を設定
S156
得点Pに0を設定
S157
更新頻度Sを算出
S158
更新頻度Sに0を設定
S159
更新頻度Sを更新
S160
くりかえし
END

【図９】

	比較結果	得点P	前回の更新頻度S_0	更新頻度S
領域１	同じ	100	73	78
領域２	同じ	100	73	78
領域３	違う	0	46	37

α＝0.8とすると

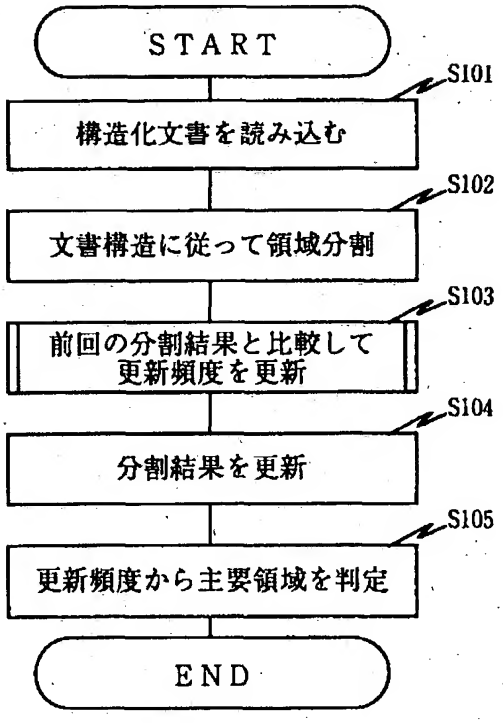
$73\times0.8 + 100\times(1-0.8) \fallingdotseq 78$

$46\times0.8 + 0\times(1-0.8) \fallingdotseq 37$

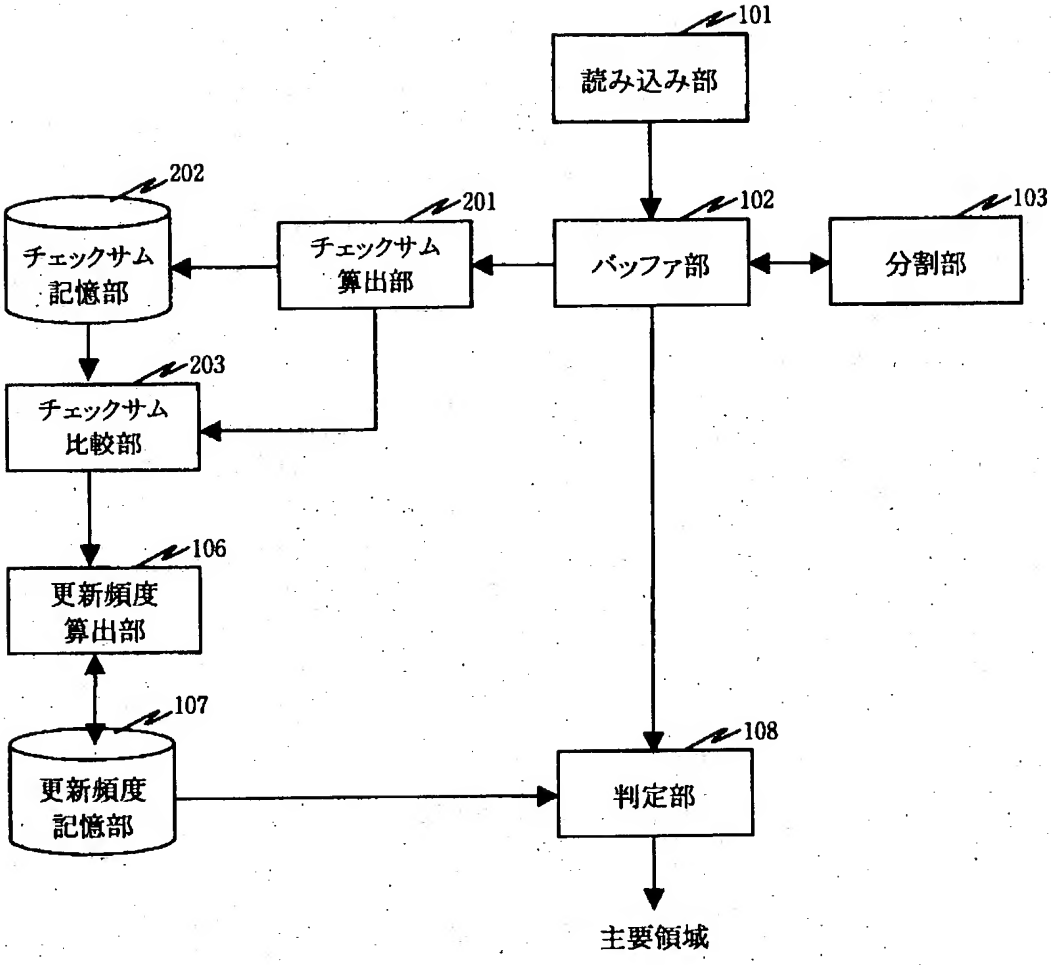
【図１０】

	比較結果	得点P	前回の更新頻度S_0	更新頻度S
領域１	同じ	100	73	78
領域２	違う	-	73	0
領域３	違う	-	46	0
領域４	比較対象なし	-	-	0

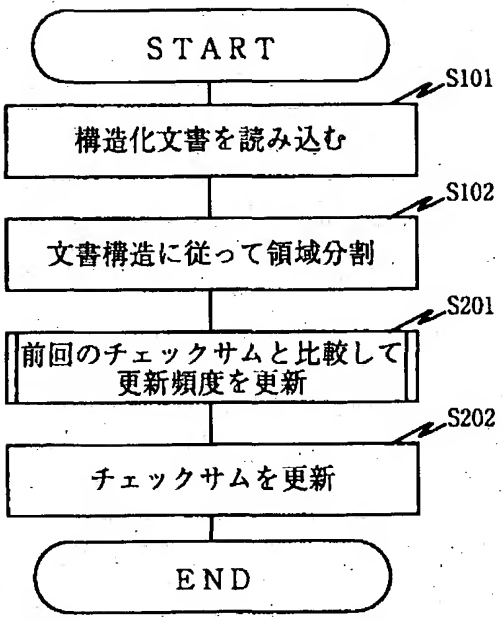
【図１１】



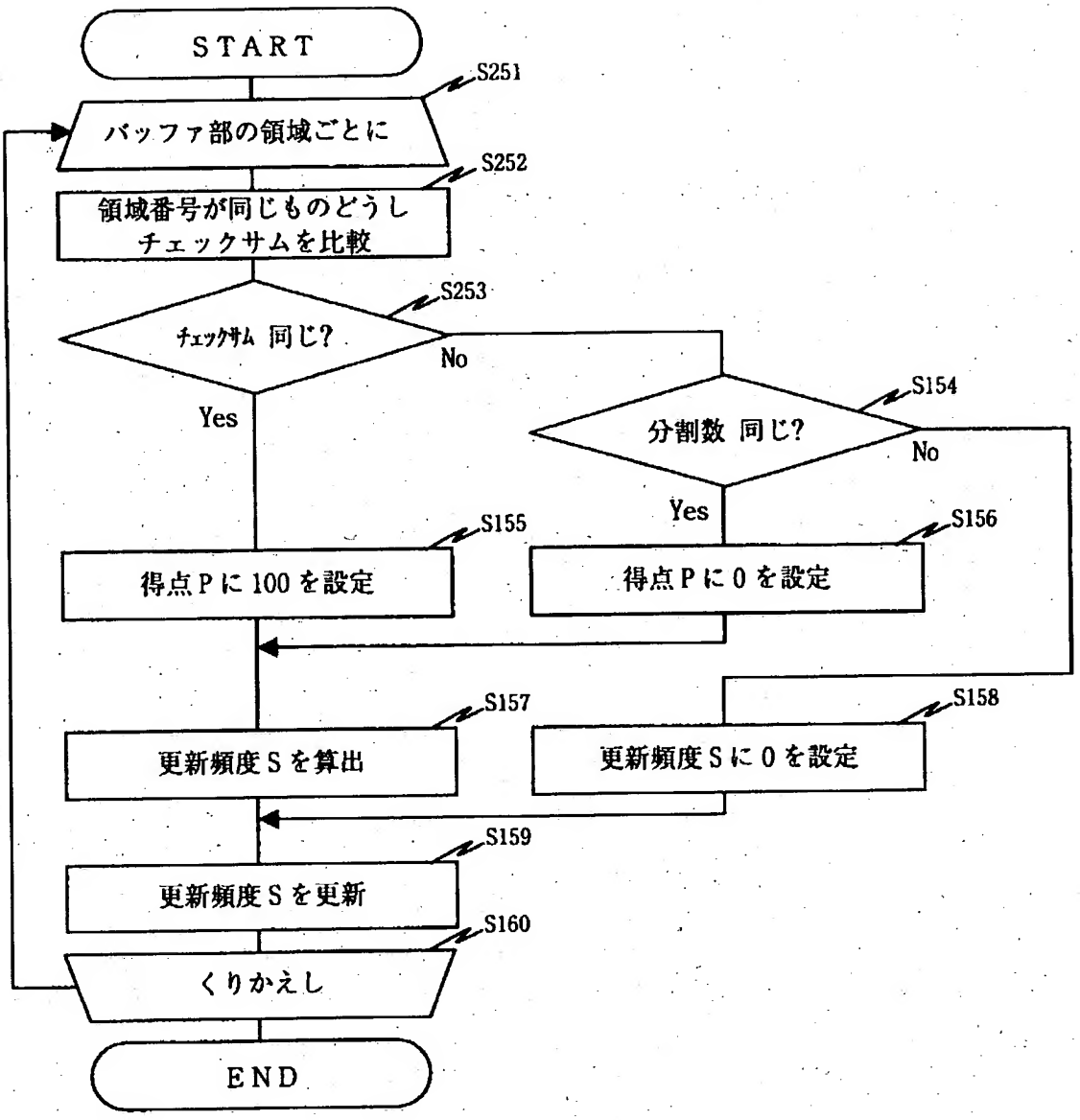
【図１２】


101
読み込み部
202
チェックサム
記憶部
201
チェックサム
算出部
102
バッファ部
103
分割部
203
チェックサム
比較部
106
更新頻度
算出部
107
更新頻度
記憶部
108
判定部
主要領域

【図１３】


START
S101
構造化文書を読み込む
S102
文書構造に従って領域分割
S201
前回のチェックサムと比較して
更新頻度を更新
S202
チェックサムを更新
END

【図１４】


START
S251
バッファ部の領域ごとに
S252
領域番号が同じものどうし
チェックサムを比較
S253
ﾁｪｯｸｻﾑ 同じ?
No
Yes
S154
分割数 同じ?
No
Yes
S155
得点Pに100を設定
S156
得点Pに0を設定
S157
更新頻度Sを算出
S158
更新頻度Sに0を設定
S159
更新頻度Sを更新
S160
くりかえし
END

【書類名】　　要約書

【要約】

【課題】　構造化文書における主要領域を判定する

【解決手段】　本発明の主要領域判定装置は、定期的または不定期に構造化文書を取得する読み込み部と、読み込んだ構造化文書を１つ以上の領域に分割する分割部と、分割結果を一時的に記憶する分割結果記憶部と、領域ごとに前回の内容と今回の内容を比較して更新有無を調べる比較部と、領域ごとに過去の更新情報を記憶する更新情報記憶部と、前回までの更新情報と今回の更新有無から更新頻度を算出する更新頻度算出部と、更新頻度が最も高い領域を主要領域として判定する判定部とを備えることを特徴する。

【選択図】　　図２

出願人履歴情報

識別番号　〔０００００２９５〕

１．変更年月日　１９９０年　８月２２日
［変更理由］　新規登録
住　所　東京都港区虎ノ門１丁目７番１２号
氏　名　沖電気工業株式会社